Biometrics 62, 478–487
June 2006

DOI: 10.1111/j.1541-0420.2005.00488.x

Effect of Misreported Family History on Mendelian Mutation Prediction Models

Hormuzd A. Katki

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, U.S.A.
Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, Room 8044, Rockville, Maryland 20852, U.S.A.
email: katkih@mail.nih.gov

Summary. People with familial history of disease often consult with genetic counselors about their chance of carrying mutations that increase disease risk. To aid them, genetic counselors use Mendelian models that predict whether the person carries deleterious mutations based on their reported family history. Such models rely on accurate reporting of each member's diagnosis and age of diagnosis, but this information may be inaccurate. Commonly encountered errors in family history can significantly distort predictions, and thus can alter the clinical management of people undergoing counseling, screening, or genetic testing. We derive general results about the distortion in the carrier probability estimate caused by misreported diagnoses in relatives. We show that the Bayes factor that channels all family history information has a convenient and intuitive interpretation. We focus on the ratio of the carrier odds given correct diagnosis versus given misreported diagnosis to measure the impact of errors. We derive the general form of this ratio and approximate it in realistic cases. Misreported age of diagnosis usually causes less distortion than misreported diagnosis. This is the first systematic quantitative assessment of the effect of misreported family history on mutation prediction. We apply the results to the BRCAPRO model, which predicts the risk of carrying a mutation in the breast and ovarian cancer genes BRCA1 and BRCA2.

Key words: Bayes factor; BRCA1; BRCA2; BRCAPRO; CRCAPRO; Family history; Genetic counseling; Measurement error.

1. Introduction

People who are concerned that their family has a high prevalence of disease may seek genetic counseling to assess their risk of carrying inherited genetic mutations that cause the disease (Croyle and Lerman, 1999). To aid such people (consultands), genetic counselors and other medical professionals (genetic counselors) employ statistical methods that predict whether the consultand carries deleterious mutations by using the consultand's reported family history of disease. For syndromes whose onset occurs over a lifetime (as is common in cancer), family history is the age at which each family member developed disease, or that member's current age or age-at-death. Mendelian models use Mendel's laws and Bayes' rule to combine family history information with each mutation's known prevalence and penetrance (probability of disease given mutations) to determine the probability that the consultand is a mutation carrier (Murphy and Mutalik, 1969). The consultand's carrier probability is a crucial component in the consultand's decision to take a genetic test (if a test exists), to undergo frequent disease screening, or to consider prophylactic options. For the breast–ovarian cancer syndrome (Claus et al., 1996), the popular Mendelian model BRCAPRO estimates the probability that a consultand carries a deleterious mutation in the BRCA1 and BRCA2 genes, based on family history of breast and ovarian cancer (see Berry et al., 1997; Parmigiani, Berry, and Aguilar, 1998). Another example is CRCAPRO, which computes the probability of carrying a mutation in the genes MLH1 and MSH2 given family history of colorectal and endometrial cancer (Chen et al., 2004).

However, Mendelian models rely on accurate knowledge of family history. Consultands cannot always provide accurate information; sometimes they cannot provide the required information, or mistakenly provide inaccurate information. Table 1 shows error rates in reporting relatives' type of cancer diagnoses (i.e., whether the relative has that type of cancer or not), by degree of relationship to the consultand. Note that 56% of ovarian cancers in grandmothers are unreported, although since ovarian cancer is so rare, 97% of grandmothers who are reported without ovarian cancer truly do not have it. Although this false positive rate is only 2%, grandmothers reported with ovarian cancer only have a 63% chance of truly having it. Reporting error rates in colorectal and endometrial cancers, which are relevant to CRCAPRO, are alarming.

Such errors can seriously distort the carrier probability estimate. For example, consider the family tree of Figure 1. If the consultand is unaware of her grandmother having ovarian cancer, by reporting her as dead by another cause, then her family only has older breast cancers. This is not strongly indicative

Table 1

Error rates (%) in reported cancer diagnoses for relatives from Ziogas and Anton-Culver (2003), Tables 3 and 4. They define 1° relatives as the consultand's parents, children, or siblings, and 2° relatives as the consultand's grandparents, aunts, and uncles. Breast and ovary are relevant for BRCAPRO. Colorectum and endometrium are relevant for CRCAPRO.

Site	Relationship	False negative rate	False positive rate	Positive predictive value	Negative predictive value
Breast	1°	5	3	89	99
Breast	2°	18	3	89	96
Ovary	1°	17	1	76	99
Ovary	2°	56	2	63	97
Colorectum	1°	10	3	80	99
Colorectum	2°	42	2	74	97
Endometrium	1°	44	2	37	99
Endometrium	2°	63	2	21	99

of any BRCA mutation in the family, and the BRCAPRO carrier probability estimate for the consultand is only 4%. But ovarian cancer at any age is a strong indicator, and the consultand's BRCAPRO probability for any BRCA mutation including this correct information jumps to 20%. This difference is especially critical because many genetic counselors offer genetic testing to the consultand once the probability exceeds 10% (Domchek et al., 2003). Also, health insurers may not cover the expense of the test unless the probability is high enough (Zielinski, 2005).

Inaccurate family history is a reality of genetic counseling. Although genetic counselors working in academia or controlled studies try to contact other relatives to verify reported family history, relatives may be deceased or otherwise difficult to contact. Furthermore, genetic counselors not working in those settings may have limited options for verifying family history. We are unaware of any methodological contribution addressing this complex problem. Although there are other types of errors, this article focuses on errors in reporting a relative's diagnosis (i.e., whether they are affected or not) and in reporting the age of that diagnosis:

1. Diagnosis incorrect, age-at-diagnosis correct: This error results from the consultand knowing that something happened to their relative at that age, but not knowing what. In the example above, ovarian cancer diagnoses are

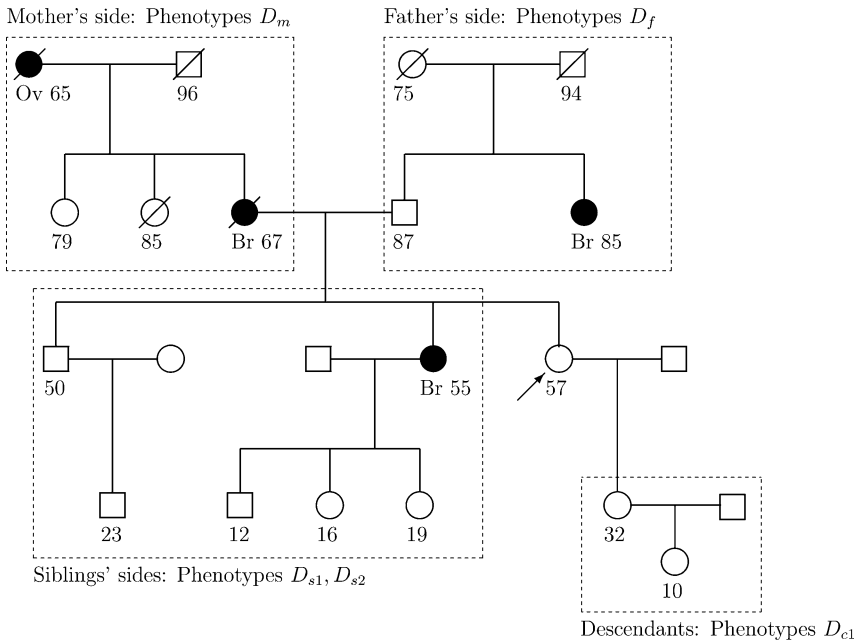

Figure 1. Family tree with breast (Br) and ovarian (Ov) cancer history. The arrow points to the consultand. Circles are females, squares are males. Slash means the relative died, dark shape means the relative got cancer, light shape and no slash means the relative is alive with no cancer, and the age of those outcomes is below each member. D_f, D_m are all the phenotypes on the father's and mother's sides, respectively. D_{c1} are the phenotypes of the sole child and grandchild. D_{s1} are the phenotypes of the 50-year-old brother and his descendants and D_{s2} are the phenotypes of the 55-year-old sister and her descendants.

often concealed from the family and reported as death by other causes. But since ovarian cancer is often rapidly fatal, then at least the age-at-death approximates the true age-at-diagnosis of ovarian cancer. This article primarily concerns this error.

2. Diagnosis correct, age-at-diagnosis incorrect: Here the consultand remembers a relative's disease diagnosis but not their age-at-diagnosis. This error also happens if the model requires age-of-onset rather than age-of-diagnosis. Often, all that is known is an age range within which the relative developed disease.
3. Diagnosis incorrect, age-at-diagnosis incorrect.

First, we derive a convenient expression for the Bayes factor (BF) from Mendelian models. Defining the distortion of the estimated carrier odds caused by errors as the ratio of the BF given correct information to the BF given misreported information, we derive the distortion caused by the first type of error in each family member. We give exact expressions, and also useful approximations for a rare autosomal-dominant mutation that are valid in general Mendelian models. The distortion will take the form of a hazard ratio. We focus on misreported diagnoses since they usually cause more distortion than misreported ages. However, we show the distortion caused by all three types of error for BRCAPRO.

2. Computing the Carrier Probability

Mendelian models require knowledge of which disease each relative developed and the age when it was diagnosed, or if no disease developed, the current age or age of death. For simplicity, assume that only one disease is involved in the syndrome, so $c_i = 0, 1$ indicates disease in relative i (let $i = 0$ be the consultand) and y_i is the age of diagnosis (or current age or age-at-death). Let $H_i = (y_i, c_i)$ and $H = (H_0, H_1, \ldots)$. Assume that individuals independently inherit one allele from each parent at each autosomal locus and that the alleles are either normal or mutated. Let $\gamma_i = 0, 1$ indicate carrying the genetic variants that confer disease risk; for example, $\gamma_i = 1$ for a dominant trait when the member carries at least one mutant allele, but for a recessive trait $\gamma_i = 1$ implies the relative carries two mutant alleles. We call γ_i the carrier status. The absolutely continuous penetrance survivals $S_0(y)$, $S_1(y)$ are the probability of surviving the disease up to age y given $\gamma_i = 0, 1$, respectively. Similarly, the penetrance densities $f_0(y)$, $f_1(y)$ are the density of getting the disease at age y given $\gamma_i = 0, 1$, respectively. The penetrance hazards are $h_0(y) = f_0(y)/S_0(y)$ and $h_1(y) = f_1(y)/S_1(y)$. The population prevalence of $\gamma_i = 1$ is π. The aim is the consultand's carrier probability $P(\gamma_0 = 1 \,|\, H)$.

By Bayes rule, the odds of the consultand being a carrier is a product of the prior odds in the population and the BF:

$$\frac{P(\gamma_0 = 1 \,|\, H)}{P(\gamma_0 = 0 \,|\, H)} = \frac{\pi}{1-\pi} \times \mathrm{BF}(H), \quad \text{where}$$

$$\mathrm{BF}(H) = \frac{P(H \,|\, \gamma_0 = 1)}{P(H \,|\, \gamma_0 = 0)}.$$

Since family history only affects the carrier probability through the BF, it suffices to consider the effect of errors in family history on the BF. Section 2.1 shows and interprets the BF, and later sections examine the effect of errors on the BF.

2.1 *Bayes Factor for a General Family*

We require two assumptions to compute the BF. The first assumption assumes that H_i is independent of all other information, given that member's γ_i. This assumption is unrealistic when, for example, other disease risk factors confer similar risk as the mutation. However, when the mutation confers such high disease risk (as BRCA mutations do) that it overwhelms the effects of other risk factors, then the assumption can be reasonable, and thus most Mendelian models are used only for highly penetrant mutations. Under this assumption, contributions to the BF from family members are channeled through the first-degree (1°) relatives (parents and children), that is, other relatives yield information about the consultand's carrier status only via the information they yield about the carrier status of 1° relatives (see Lauritzen and Sheehan, 2003 for more on family tree likelihood calculations).

The second assumption is no inbreeding between family members. In particular, this means no inbreeding within or between ancestors and descendants. Relative to the consultand, any family is composed of "ancestors" (the father's and mother's sides, as well as siblings and their families) and "descendants" (children of the consultand and their families); see Figure 1. No inbreeding within ancestors or descendants means that conditioning on the 1° relatives makes their families contribute independently to the likelihood (Lauritzen and Sheehan, 2003). No inbreeding between the ancestors and descendants implies that all relationships between the consultand's ancestors and descendants are via the consultand, as in Figure 1. For example, this would not be true if the 32-year-old and 23-year-old cousins in Figure 1 had a child. The family tree is a Bayesian network where, because of the no-inbreeding condition and the first assumption, the consultand's carrier status γ_0 is the only conduit of information between the ancestors and descendants; thus conditioning the likelihood on γ_0 makes the ancestors independent of the descendants (Lauritzen and Sheehan, 2003). Thus the likelihood $P(H \,|\, \gamma_0)$ has independent contributions from the (c)onsultand, the (a)ncestors, and the (d)escendants, so $\mathrm{BF}(H) = \mathrm{BF_c} \times \mathrm{BF_a} \times \mathrm{BF_d}$.

To calculate the consultand contribution $\mathrm{BF_c}$, the likelihood is

$$P(H_0 \,|\, \gamma_0) = P(Y_0 = y_0, C_0 = c_0 \,|\, \gamma_0) \propto f_{\gamma_0}(y_0)^{c_0} S_{\gamma_0}(y_0)^{1-c_0},$$

assuming independent noninformative censoring. It can be shown that this censoring mechanism drops out of the likelihood (H. A. Katki, S. Chen, and G. Parmigiani, unpublished manuscript) and will not be explicitly referred to. The consultand's contribution to the BF is

$$\mathrm{BF_c} = \frac{P(H_0 \,|\, \gamma_0 = 1)}{P(H_0 \,|\, \gamma_0 = 0)} = \frac{f_1(y_0)^{c_0} S_1(y_0)^{1-c_0}}{f_0(y_0)^{c_0} S_0(y_0)^{1-c_0}}. \qquad (1)$$

$\mathrm{BF_c}$ is a ratio of penetrance densities if the consultand has the disease, otherwise, it is the ratio of penetrance survivals.

We calculate $\mathrm{BF_d}$ by conditioning on the children's carrier status, which breaks the descendants into independent

contributions from each of K children and their descendants. Define each D_{ck}, $k = 1, \ldots, K$ as the collection of the phenotypes H within the family of each child, for example, in Figure 1, D_{c1} are the phenotypes of the consultand's sole child and grandchild. The descendants' likelihood contribution is

$$P(D_{c1}, \ldots, D_{cK} \,|\, \gamma_0) = \prod_{k=1}^{K} \sum_{i=0}^{1} P(D_{ck} \,|\, \gamma_k = i) P(\gamma_k = i \,|\, \gamma_0) = \prod_{k=1}^{K} \sum_{i=0}^{1} L_{ik}(\gamma_0),$$

where γ_k remains the carrier status of child k, and without loss of generality, only considers i such that $L_{ik}(0) \neq 0$. Then, $W = \prod_{i=0}^{1} 1/L_{ik}(0)$ and weights $w_i(k) = L_{ik}(0) \times W$ are well defined. Then, the descendants' BF contribution is

$$\mathrm{BF_d} = \prod_{k=1}^{K} \frac{\sum_{i=0}^{1} L_{ik}(1)}{\sum_{i=0}^{1} L_{ik}(0)} = \prod_{k=1}^{K} \frac{\sum_{i=0}^{1} L_{ik}(1) \times W}{\sum_{i=0}^{1} L_{ik}(0) \times W} = \prod_{k=1}^{K} \frac{w_0(k) B_0 + w_1(k) B_0}{w_0(k) + w_1(k)},$$

where $B_i = P(\gamma_k = i \,|\, \gamma_0 = 1)/P(\gamma_k = i \,|\, \gamma_0 = 0)$ are the BFs favoring the consultand being a carrier, if the child's carrier status is known. Thus $\mathrm{BF_d}$ has the appealing interpretation of a product of weighted averages of the BFs that each child contributes if their carrier status were known. Since their carrier status is unknown, the BFs are averaged with weights reflecting the likelihood of each child's possible carrier status.

To calculate $\mathrm{BF_a}$, the 1° relatives are the parents, so conditioning the likelihood on parental carrier status breaks the ancestors into independent contributions from mother's side ancestors, father's side ancestors, and siblings' families (who contribute only through their connection to the parents); see Figure 1. Let D_f be the collection of all the phenotypes on the father's side, γ_f be the father's carrier status, D_m be the collection of all mother's side phenotypes, γ_m be the mother's carrier status, and each $D_{sk}, k = 1, \ldots, S$ be the collection of all the phenotypes on each of the S sibling's sides. The ancestors' likelihood contribution $P(D_f, D_m, D_s \,|\, \gamma_0)$ is

$$\sum_{i,j=0}^{1} \underbrace{\left(\prod_{k=1}^{S} P(D_{sk} \,|\, \gamma_f = i, \gamma_m = j) \right) P(D_f \,|\, \gamma_f = i) P(D_m \,|\, \gamma_m = j)}_{a_{ij}} \times \underbrace{P(\gamma_f = i, \gamma_m = j \,|\, \gamma_0)}_{b_{ij}(\gamma_0)}. \quad (2)$$

Without loss of generality, only consider i such that $a_{ij} b_{ij}(0) \neq 0$. Then, $W = 1/\prod_{i,j=0}^{1} a_{ij} b_{ij}(0)$ and weights $w_{ij} = a_{ij} b_{ij}(0) \times W$ are well defined. Then the ancestors' BF contribution is

$$\mathrm{BF_a} = \frac{\sum_{i,j=0}^{1} a_{ij} b_{ij}(1)}{\sum_{i,j=0}^{1} a_{ij} b_{ij}(0)} = \frac{\sum_{i,j=0}^{1} a_{ij} b_{ij}(1) \times W}{\sum_{i,j=0}^{1} a_{ij} b_{ij}(0) \times W} = \frac{w_{00} B_{00} + w_{01} B_{01} + w_{10} B_{10} + w_{11} B_{11}}{w_{00} + w_{01} + w_{10} + w_{11}},$$

and B_{ij} is the BF contributed by the parents if their carrier status were known:

$$B_{ij} = \frac{b(1)_{ij}}{b(0)_{ij}} = \frac{P(\gamma_f = i, \gamma_m = j \,|\, \gamma_0 = 1)}{P(\gamma_f = i, \gamma_m = j \,|\, \gamma_0 = 0)}.$$

Thus, similar in interpretation to $\mathrm{BF_d}$, $\mathrm{BF_a}$ is a weighted average of BFs that the parents would contribute if their carrier status were known, with weights reflecting the likelihood of the parents' possible carrier status.

Thus, under the assumptions outlined above, the BF breaks into contributions from the consultand, the ancestors, and the descendants. The ancestors and descendants contribute a weighted average of the BFs from 1° relatives over their possible carrier status, weighted by the likelihood of each possible carrier status. The Appendix details these BF calculations for the simple families of consultand/1° relative and consultand/1° relative/2° relative, which we use through the rest of the article. The Appendix also provides approximate BFs for a rare autosomal-dominant mutation. The rest of the article shows the distortion of BF by the three types of errors of Section 1.

3. Effect of Misreported Diagnosis in Each Relative

Under the first type of error (which assumes known age of diagnosis), the effect of misreported diagnosis on the BF can be summarized by a ratio of BFs, the numerator being the BF for the family member correctly having disease and the denominator for the family member incorrectly not having disease. This is an underreporting error; for overreporting error, the ratio is inverted.

If the consultand has underreported disease then the BF ratio is

$$\text{BF ratio} = \frac{f_1/f_0}{S_1/S_0} = \frac{f_1/S_1}{f_0/S_0} = \frac{h_1}{h_0}, \quad (3)$$

the hazard ratio of the mutation versus wild-type (since age of diagnosis is known, it is dropped as an argument). Thus errors in diagnosis are irrelevant if the diagnoses imply equivalent hazards at the age of diagnosis. Furthermore, if the hazards are proportional, then misreporting causes the same distortion at any age. Since the BF factors into contributions from the consultand and from the family, the BF ratio for the misreporting in the family is a product of (3) with the BF ratio from the family derived below. It is most useful to consider the effect of misreporting diagnosis in a single 1° or 2° relative.

3.1 *Misreporting a 1° Relative's Diagnosis*

In a consultand/1° relative family with underreported 1° relative's diagnosis, then the ratio of BFs is (see the Appendix for derivation of the BF)

$$\text{BF ratio} = \frac{f_0 p_{10} + f_1 p_{11}}{f_0 p_{00} + f_1 p_{01}} \times \frac{S_0 p_{00} + S_1 p_{01}}{S_0 p_{10} + S_1 p_{11}} = \frac{h_1^{(1)}}{h_0^{(1)}}, \tag{4}$$

$$h_j^{(1)} = \frac{(S_1 p_{j1})^{-1} h_0 + (S_0 p_{j0})^{-1} h_1}{(S_1 p_{j1})^{-1} + (S_0 p_{j0})^{-1}}, \tag{5}$$

where $p_{ij} = P(\gamma_1 = j \,|\, \gamma_0 = i)$ and $S_j p_{ij} = P(Y_1 = y_1, C_1 = 0, \gamma_1 = j \,|\, \gamma_0 = i)$. The $h_j^{(1)}$ are weighted averages of h_0, h_1. Since a ratio of weighted averages of two quantities is less than the ratio of the two quantities, misreporting a 1° relative distorts the BF less than the same misreport in the consultand. Furthermore, the weighting depends on $S_j p_{ij}$, the likelihood of the 1° relative surviving, given the consultand. If this probability for j is larger, then h_j gets more weight. If the hazards are proportional, there is still time dependence in equation (4) since the weights depend on time through the survivals.

For a rare autosomal-dominant mutation (see equation (A.5)), the BF ratio is approximately

$$\begin{aligned}\frac{h_1^{(1)}}{h_0^{(1)}} &\approx \frac{1 + \dfrac{f_1}{f_0}}{1 + \dfrac{S_1}{S_0}} = \frac{1}{S_0 + S_1}\left(S_0 + S_1 \frac{h_1}{h_0}\right) \\ &= \frac{1}{1 + \dfrac{S_1}{S_0}}\left(1 + \frac{S_1}{S_0}\frac{h_1}{h_0}\right). \end{aligned} \tag{6}$$

Equations (6) and (3) are the same except for the "1+" terms, and (6) is a linear function of the hazard ratio, which is the effect of an underreport for the consultand. The intercept and slope sum to one, and each are posterior probabilities of being a noncarrier (and carrier, respectively) given being unaffected, assuming even prior odds of being a carrier. Thus, when $h_1/h_0 = 1$, then $h_1^{(1)}/h_0^{(1)} = 1$; no distortion in the consultand is also no distortion for the 1° relative. The distortion caused by misreported diagnosis in 1° relatives is less than that caused by the same misreported diagnosis in the consultand (Figure 2). At early ages of diagnosis, when $S_0 \approx S_1 \approx 1$, the distortion in the BF by misreported diagnoses in 1° relatives is half that of the same misreported diagnosis in the consultand. Furthermore, if the mutation is strongly deleterious and shifts the age of onset to earlier ages (like BRCA), then the survival ratio is small at older ages and the hazard ratio decreases at older ages, and the distortion attenuates. Thus misreported diagnoses in old enough 1° relatives cause little distortion. Intuitively, if $S_1(y) \approx 0$, then the relative has reached an age that carriers rarely reach, thus the relative is unlikely to be a carrier. It does not matter what happens at age y, just reaching that age suffices.

3.2 *Misreporting a 2° Relative's Diagnosis*

For the consultand/1° relative/2° relative family (see Appendix equation (A.6)), the BF ratios, depending on the 1° relative's status c_1, are $h_1^{(2)}/h_0^{(2)}$ where

$$h_i^{(2)} = \frac{(f_1^{c_1}(y_1) S_1^{1-c_1}(y_1) S_1(y_2) p_{i1})^{-1} h_0(y_2) + (f_0^{c_1}(y_1) S_0^{1-c_1}(y_1) S_0(y_2) p_{i0})^{-1} h_1(y_2)}{(f_1^{c_1}(y_1) S_1^{1-c_1}(y_1) S_1(y_2) p_{i1})^{-1} + (f_0^{c_1}(y_1) S_0^{1-c_1}(y_1) S_0(y_2) p_{i0})^{-1}}.$$

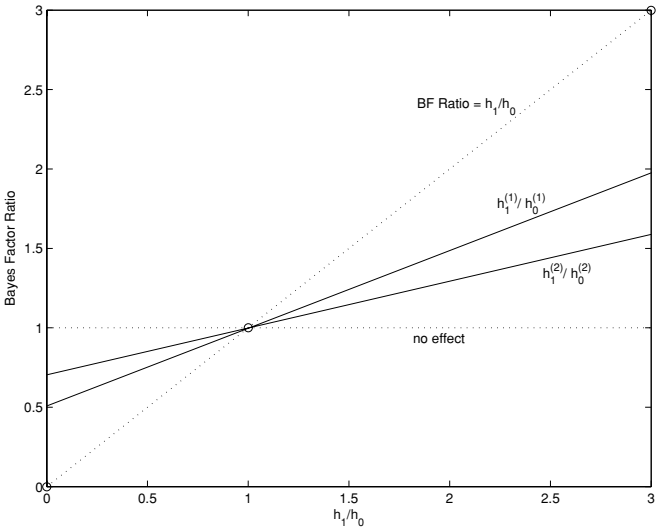

Figure 2. Bayes factor ratio in 1° and 2° relatives as a function of the BF ratio in the consultand (which is the hazard ratio).

Again, this is a ratio of weighted averages of the hazards. Under the rare autosomal-dominant mutation approximation of (A.7), the ratios are

$$\begin{aligned}\text{BF}^1_{\text{ratio}} &= \frac{\text{BF}(c_1 = 1, c_2 = 1)}{\text{BF}(c_1 = 1, c_2 = 0)} \\ &\approx \frac{1 + \dfrac{f_1(y_1)}{2 f_0(y_1)}\left(1 + \dfrac{f_1(y_2)}{f_0(y_2)}\right)}{1 + \dfrac{f_1(y_1)}{2 f_0(y_1)}\left(1 + \dfrac{S_1(y_2)}{S_0(y_2)}\right)}, \end{aligned} \tag{7}$$

$$\begin{aligned}\text{BF}^0_{\text{ratio}} &= \frac{\text{BF}(c_1 = 0, c_2 = 1)}{\text{BF}(c_1 = 0, c_2 = 0)} \\ &\approx \frac{1 + \dfrac{S_1(y_1)}{2 S_0(y_1)}\left(1 + \dfrac{f_1(y_2)}{f_0(y_2)}\right)}{1 + \dfrac{S_1(y_1)}{2 S_0(y_1)}\left(1 + \dfrac{S_1(y_2)}{S_0(y_2)}\right)}. \end{aligned} \tag{8}$$

These ratios are of the form $g(\alpha_1) = (1 + \alpha_1(1 + A))/(1 + \alpha_1(1 + B))$ where α_1 is contributed by the 1° relative and $(1 + A)/(1 + B)$ is the ratio if the 1° relative were misreported (equation (6)). Since $1 \le g(\alpha_1) \le (1 + A)/(1 + B)$, the effect of misreported diagnosis in the 2° relative is attenuated from the effect that the same error in the 1° relative would have had.

In the below derivations, the density ratios and survival ratios are all for 2° relative's time y_2, so the time arguments are suppressed. Note that

$$\begin{aligned}\frac{h_1^{(2)}}{h_0^{(2)}} &\approx \frac{1 + \alpha_1\left(1 + \dfrac{f_1}{f_0}\right)}{1 + \alpha_1\left(1 + \dfrac{S_1}{S_0}\right)} \\ &= \frac{1}{1 + \alpha_1\left(1 + \dfrac{S_1}{S_0}\right)} \times \alpha_1\left(1 + \frac{S_1}{S_0}\right)\frac{h_1^{(1)}}{h_0^{(1)}}, \end{aligned}$$

a linear function of the BF ratio for the 1° relative. The intercept and slope are less than one, so we again have attenuation. Plugging in equation (6),

$$\frac{h_1^{(2)}}{h_0^{(2)}} \approx \frac{1}{1+\alpha_1\left(1+\dfrac{S_1}{S_0}\right)} \times \left(1+\alpha_1+\alpha_1\frac{S_1}{S_0}\frac{h_1}{h_0}\right),$$

the BF ratio in the 2° relative is linear in the BF ratio for the consultand (Figure 2). Thus, since $g(\alpha_1)$ is monotone increasing, and if $f_1/f_0 > S_1/S_0$ (but small enough so the rare mutation approximation [A.7] still holds; reasonable for BRCA after age 35), then misreported diagnosis in the 2° relative causes greater distortion if the 1° relative is affected rather than unaffected. But the change in the distortion with age of the 2° relative strongly depends on the status and age of the 1° relative, so no simple rules can be stated.

3.3 *Application to BRCAPRO*

Figure 3 shows the distortion caused by misreported diagnosis in BRCAPRO, with penetrances from S. Chen et al. (unpublished manuscript). Underreporting in consultands is the hazard ratio of (3). The worst error is underreporting ovarian cancer in consultands, because ovarian cancer yields higher penetrance density ratios than breast cancer, especially at older ages. In fact, underreported ovarian cancer in 1° relatives is worse than underreported breast cancer in consultands. The weakest errors all involve underreporting breast cancer in 2° relatives.

The situation is interesting for underreported ovarian cancer in 2° relatives: At younger ages, the error is worst if 1° relatives have breast cancer, at older ages if they have ovarian cancer. Most interestingly, at young ages, the error is worse if 1° relatives have no cancer than if they have ovarian cancer.

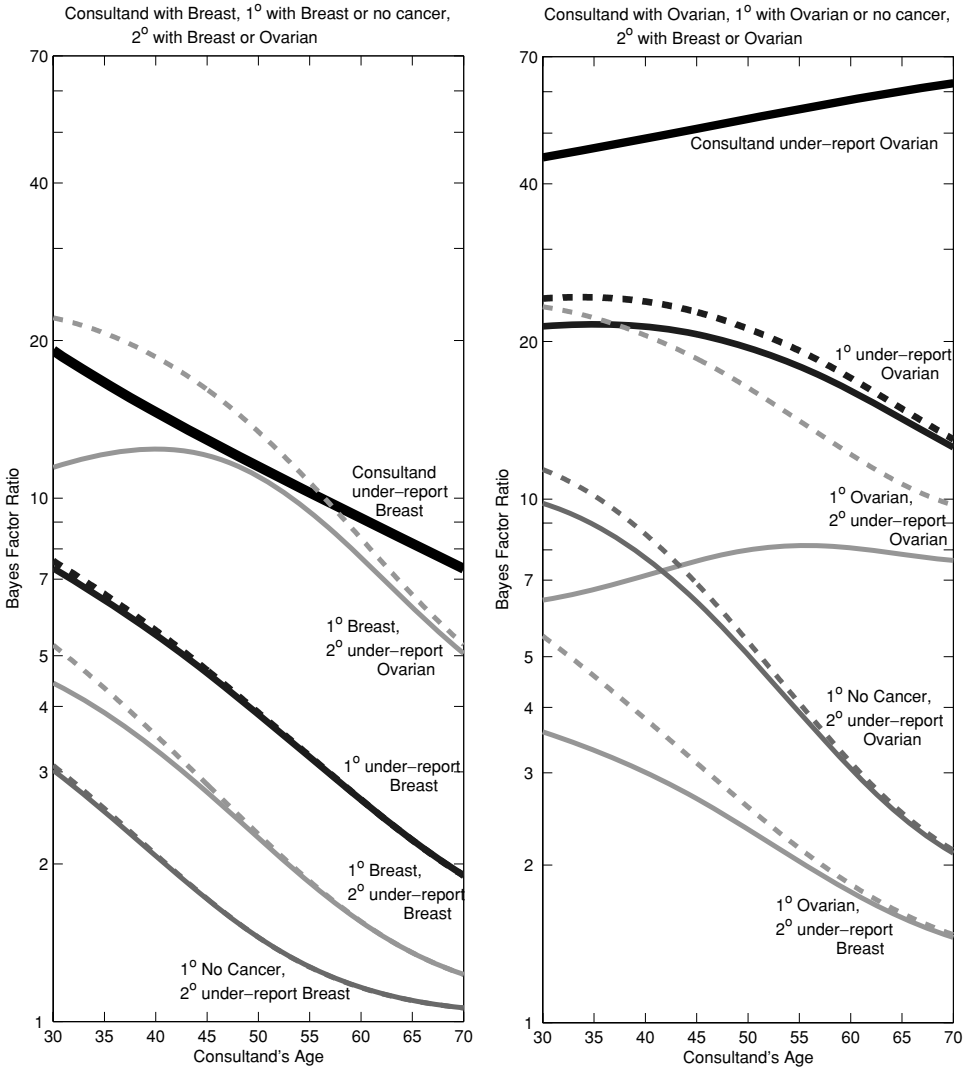

Figure 3. Distortion of the Bayes factor for BRCA1 carrier probability from BRCAPRO by underreporting breast and ovarian cancer diagnoses in the consultand, 1°, and 2° relatives (for consultand 10, 20 years younger than 1°, 2° relatives, respectively). The ten situations are split among two graphs to minimize overlap. Solid lines are the exact BF ratio for each situation, nearest dashed line is the corresponding rare mutation approximation. The darkest lines are for consultand-only or 1°-only families, gray lines are families with 1° and 2° relatives. Penetrances are from S. Chen et al. (unpublished manuscript).

This contradicts the standard intuition that because cancers in each generation on a side of the family are indicative of a mutation within the family, an error on that side of a family is more devastating. Figure 3 shows that even if the 1° relative has no cancer, 2° relatives underreporting ovarian cancer remains an important error.

The example family of Figure 1 is more complex than those represented in this figure, but is closest to the 1° Breast 2° underreport Ovarian line. For this family, being unaware of the grandmother's ovarian cancer leads to a BRCAPRO probability for a BRCA1 mutation of 1.6%, while the probability based on correct information is 10%, for a BF ratio of 6.8.

The approximations for 1° and 2° relatives overestimate the effect of underreporting and are rougher when underreporting ovarian cancer because the penetrance density ratio is higher for ovarian cancer than breast cancer, especially at younger ages. Also, the requirements of the approximation for 2° relatives are more stringent than for 1° relatives (see the Appendix).

Any of these errors, especially underreporting ovarian cancer in any relative, can cause major distortion. Since BRCA mutations are rare, an underreporting BF ratio of just two can halve the carrier probability. Many of the errors are much worse.

4. Effect of Errors in Age of Diagnosis

Errors in age of diagnosis can be summarized by ratios of BFs, the numerator with the BF given the true age and denominator the BF with the incorrectly reported age. Often a consultand cannot recall the exact age, but can specify a range within which the true age lies. This is like a rounding error. Although general results are difficult to obtain, it is easy to simulate such errors in a specific model, such as BRCAPRO. For the simple families of the Appendix, the two left panels of Figure 4 show the distortion in the BF caused by rounding the age of a single 1° relative to the nearest 30 (±15 years) and 10 (±5 years) and the two right panels of Figure 4 show the effect of rounding the age of a single 2° relative to the nearest 30 and 10, assuming an affected 1° relative with known age of diagnosis at the true age of diagnosis of the 2° relative. Errors of ±15 years cause a maximum distortion of 70% in the BF, and a maximum 20% distortion for ±5 years.

Similar computations show that the BF ratio for rounding age to the nearest 30 in unaffected relatives is no more than 1.15, and the BF ratio for rounding ovarian cancer diagnosis ages to the nearest 30 in affected or unaffected relatives is no more than 1.25. These distortions are smaller because the hazards of these events are more constant over age than breast cancer.

Figure 5 shows the effect of simultaneous underreporting and rounding errors for breast cancer. Clearly, underreporting diagnosis is the more important error. For ovarian cancer, since its hazard ratio does not change much over age (see the consultand underreport ovarian line in Figure 3), errors in age of diagnosis are dominated by errors in diagnosis. However, Figure 4 shows that strong rounding error can cause meaningful distortion.

5. Discussion

The three types of misreported family history discussed in this article can seriously distort the carrier probability estimate

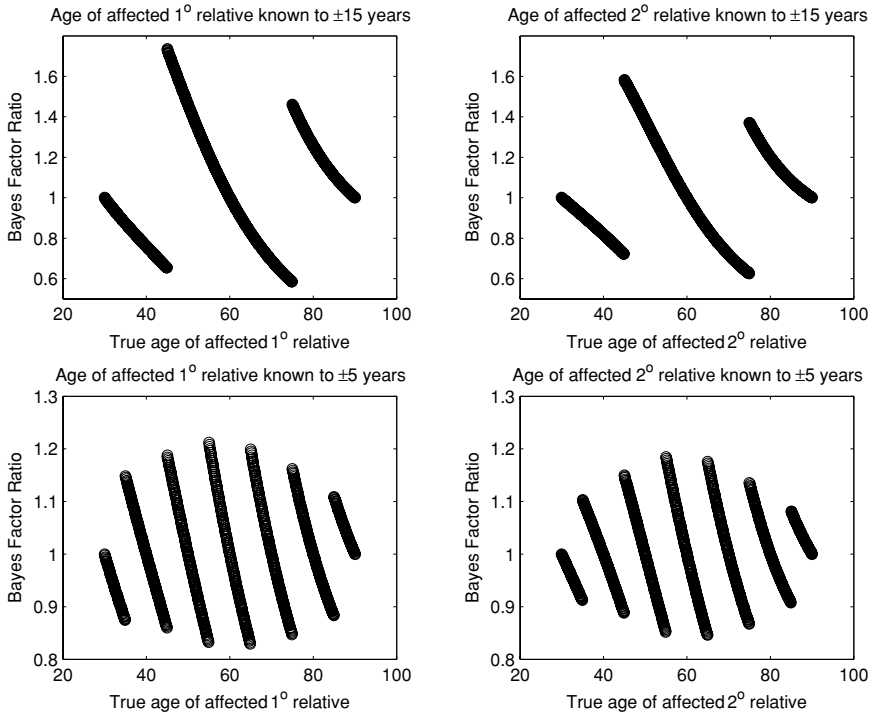

Figure 4. Distortion of the Bayes factor for BRCA1 carrier probability from BRCAPRO caused by rounding error in age of breast cancer diagnosis for affected 1°, 2° relatives. Rounding is to nearest 30 (±15 years) or nearest 10 (±5 years). The 2° relatives have the 1° relative known affected at the true age of diagnosis of the 2° relative.

from Mendelian mutation prediction models. As Table 1 shows, misreported family history is a reality of genetic counseling. The distortion caused by misreported family history is channeled through the ratio by which it changes the true BF. The BF is the product of contributions from the consultand, ancestors, and descendants that are averages of the BFs that parents and children over their possible carrier status weighted by the likelihood of each possible carrier status.

Under this convenient structure, the BF ratios for misdiagnosis in a relative (assuming known age of diagnosis) are ratios of weighted averages of the mutant and wild-type penetrance hazards. These ratios are generally complicated functions involving the entire family, but can be simplified in certain realistic cases. Assuming a rare autosomal-dominant mutation (like BRCA), the BF ratios in higher-degree relatives are linear in the BF ratio for the consultand and are attenuated toward one. In this case, the distortion in the BF by misreporting the diagnosis of 1° relatives at young ages is half that of underreporting the consultand, and the distortion attenuates at old enough ages. If the penetrance density ratio is greater than the survival ratio (like BRCA), then the effect of this error in 2° relatives is stronger if the 1° relative is affected. If the penetrance hazards are proportional (approximately true for BRCA1 and ovarian cancer), then the distortion caused by misdiagnosis in the consultand is the same for any age of diagnosis, although the distortion caused by this error in relatives generally depends on age.

These results were applied to BRCAPRO in Figures 3–5. The amount of distortion usually decreases with age of relative, but also depends on the relationship to the consultand and outcomes in other relatives. Ovarian cancer misdiagnoses at any age have strong effects for any relative. Usually, misdiagnosis in 2° relatives causes more distortion if the 1° relative is affected, but not always. Misreporting age of diagnosis in BRCAPRO for unaffected relatives has little effect. But the same error in affected relatives, while not as important as misreporting diagnosis, can have meaningful effects if the error is big enough. For a rare mutation like BRCA, an underreporting BF ratio of just two will halve the carrier probability, and Figures 3 and 5 show that the distortion is often much worse.

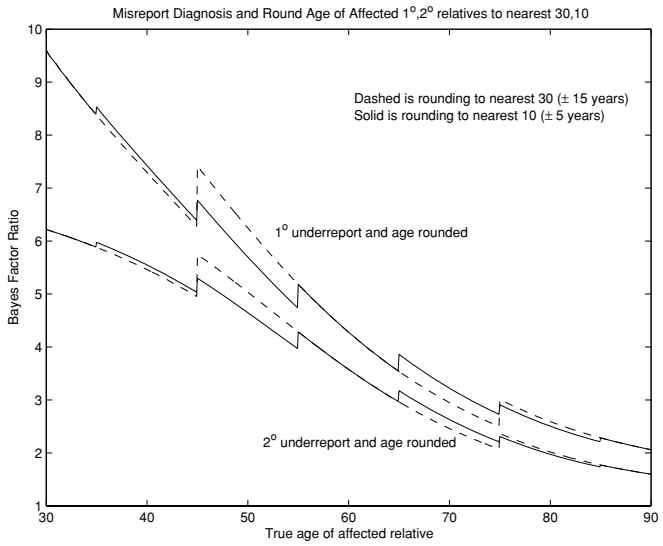

Figure 5. Effect of underreporting breast cancer diagnoses and rounding error in age of diagnosis for affected 1°, 2° relatives on the Bayes factor for BRCA1 carrier probability from BRCAPRO. The top two lines are for underreporting the 1° relative, and bottom two for the 2° relative. For the 2° relatives, they have 1° relative known affected at the true age of diagnosis of the 2° relative. Solid line is rounding to nearest 10, dashed line is nearest 30.

This article does not consider other errors in family history, such as errors in paternity. This can have enormous impact and could happen to children raised by single mothers. However, such consultands may be less likely to be sufficiently aware of their family to present for genetic counseling, and counselors will be suspicious about their knowledge of their biological father. Nevertheless, errors in paternity are an omnipresent issue.

A more common error is ignorance of the existence of certain relatives. Since consultands usually come to the clinic because of relatives who have disease, it is more common for consultands to forget about relatives who do not have disease, leading to overestimated carrier probabilities. All people have two parents and consultands usually remember their children, so all 1° relatives are usually known. For 2° relatives, all people have four grandparents, and consultands probably remember the existence of their siblings. The situation is dicier with 3° relatives such as uncles and aunts, and even worse for 4° relatives such as cousins, simply because there can be many of them. Also, error rates in diagnosis and age of diagnosis are even higher for cousins (Ziogas and Anton-Culver, 2003). Each cousin is more distant from the consultand and provides less information. Thus BRCAPRO and this article exclude cousins. However, many consultands come to the clinic because of disease clustering in their cousins, so extracting information from cousins while handling their greater reporting bias and errors in diagnosis and age of diagnosis remains an important topic.

Our results can alert genetic counselors to which types of reported family history require verification, both for general Mendelian models and BRCAPRO in particular. If a counselor suspects a single reporting error, then a sensitivity analysis is easy, and if there is a big difference, the counselor will seek permission to access the relative's medical file to verify the report.

However, counselors do not always have time to try all combinations of multiple possible errors or to verify all family members, especially relatives reported as unaffected. This article shows that diagnosis errors in 2° relatives are usually most influential if the 1° relative has disease, but not always as in BRCAPRO for ovarian cancer in young 2° relatives where the 1° relative has no cancer. Misreporting ovarian cancers in any relative at any age can cause major distortion. Errors shown to be secondary in importance are diagnosis errors in relatives known to be old enough to have reached ages that carriers rarely reach and, for BRCAPRO, errors in age of diagnosis. Errors in age of diagnosis have little effect if the penetrance hazards are flat over age. These results help counselors

to understand and prioritize the effects of errors for allocating limited resources for family history verification.

However, many genetic counselors have limited options for verification. If family history is nonverifiable and potential errors matter, then clinical decisions are complicated. In this realistic case, estimating the carrier probability requires knowing the likelihood of each possible error. Thus, the next step is extending Mendelian models to automatically handle misreported family history by accounting for population-based misreporting rates, such as those in Table 1. Such models could provide more appropriate carrier probability estimates and improve clinical management of consultands.

Acknowledgements

I greatly thank my Ph.D. thesis advisor, Giovanni Parmigiani, for his constant encouragement and guidance in this fruitful area of research, which comprises part of my thesis. I also thank my supervisor at the NCI, Barry Graubard, for his advice and support. Finally, I thank the two anonymous referees for their insightful and detailed comments. This research was supported, in part, by the Intramural Research Program of the NIH, National Cancer Institute.

References

Berry, D. A., Parmigiani, G., Sanchez, J., Schildkraut, J., and Winer, E. (1997). Probability of carrying a mutation of breast–ovarian cancer gene BRCA1 based on family history. *Journal of the National Cancer Institute* **89,** 227–238.

Chen, S., Wang, W., Broman, K. W., Katki, H. A., and Parmigiani, G. (2004). BayesMendel: An R environment for Mendelian risk prediction. *Statistical Applications in Genetics and Molecular Biology* **3**. Available at `http://www.bepress.com/sagmb/vol3/iss1/art21`.

Claus, E. B., Schildkraut, J. M., Thompson, W. D., and Risch, N. J. (1996). The genetic attributable risk of breast and ovarian cancer. *Cancer* **77,** 2318–2324.

Croyle, R. T. and Lerman, C. (1999). Risk communication in genetic testing for cancer susceptibility. *Journal of the National Cancer Institute Monographs* 59–66.

Domchek, S. M., Eisen, A., Calzone, K., Stopfer, J., Blackwood, A., and Weber, B. L. (2003). Application of breast cancer risk prediction models in clinical practice. *Journal of Clinical Oncology* **21,** 593–601.

Lauritzen, S. and Sheehan, N. (2003). Graphical models for genetic analyses. *Statistical Science* **18,** 489–514.

Murphy, E. A. and Mutalik, G. S. (1969). The application of Bayesian methods in genetic counselling. *Human Heredity* **19,** 126–151.

Parmigiani, G., Berry, D., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* **62,** 145–158.

Zielinski, S. L. (2005). As genetic tests move into the mainstream, challenges await for doctors and patients. *Journal of the National Cancer Institute* **97,** 334–336.

Ziogas, A. and Anton-Culver, H. (2003). Validation of family history data in cancer family registries. *American Journal of Preventive Medicine* **24,** 190–198.

Received September 2004. *Revised July* 2005. *Accepted August* 2005.

Appendix

Bayes Factors for Simple Families

This appendix calculates the Bayes factor (BF) for a consultand 1° relative family and consultand/1° relative/2° relative family. These concrete calculations lend intuition to the general calculations of Section 2.1. Subscripts on H, γ denote degree of relationship to consultand.

Consultand/1° Relative Family

The likelihood is

$$P(H\,|\,\gamma_0) = P(H_0, H_1\,|\,\gamma_0, \gamma_1 = 0)P(\gamma_1 = 0\,|\,\gamma_0) + P(H_0, H_1\,|\,\gamma_0, \gamma_1 = 1)P(\gamma_1 = 1\,|\,\gamma_0). \quad \text{(A.1)}$$

Assuming conditional independence of phenotypes given a member's carrier status (see Section 2.1), and ignoring independent noninformative censoring (for more, see H. A. Katki et al., unpublished manuscript), then

$$P(H_0, H_1\,|\,\gamma_0, \gamma_1) = P(H_0\,|\,\gamma_0)P(H_1\,|\,\gamma_1) = f_{\gamma_0}(y_0)^{c_0} S_{\gamma_0}(y_0)^{1-c_0} f_{\gamma_1}(y_1)^{c_1} S_{\gamma_1}(y_1)^{1-c_1}.$$

We calculate the transmission probabilities $p_{ij} = P(\gamma_1 = j\,|\,\gamma_0 = i)$ later. For now, plugging these into (A.1), the likelihood is

$$P(H\,|\,\gamma_0) = f_{\gamma_0}(y_0)^{c_0} S_{\gamma_0}(y_0)^{1-c_0} \times \big(f_0(y_1)^{c_1} S_0(y_1)^{1-c_1} p_{\gamma_0 0} + f_1(y_1)^{c_1} S_1(y_1)^{1-c_1} p_{\gamma_0 1}\big),$$

and the BF is

$$\begin{aligned} \mathrm{BF}(H) &= \frac{f_1(y_0)^{c_0} S_1(y_0)^{1-c_0}}{f_0(y_0)^{c_0} S_0(y_0)^{1-c_0}} \times \frac{f_0(y_1)^{c_1} S_0(y_1)^{1-c_1} p_{10} + f_1(y_1)^{c_1} S_1(y_1)^{1-c_1} p_{11}}{f_0(y_1)^{c_1} S_0(y_1)^{1-c_1} p_{00} + f_1(y_1)^{c_1} S_1(y_1)^{1-c_1} p_{01}} \\ &= \mathrm{BF}_{\mathrm{c}} \times \frac{M_{10} + M_{11}}{M_{00} + M_{01}} \\ &= \mathrm{BF}_{\mathrm{c}} \times \frac{\frac{p_{10}}{p_{00}}\frac{1}{M_{01}} + \frac{p_{11}}{p_{01}}\frac{1}{M_{00}}}{\frac{1}{M_{01}} + \frac{1}{M_{00}}}, \end{aligned} \quad \text{(A.2)}$$

where $M_{ij} = P(H_1\,|\,\gamma_1 = j)P(\gamma_1 = j\,|\,\gamma_0 = i)$. The BF breaks into a product of the BF from the consultand (BF_{c}) and a contribution from the 1° relative. Furthermore,

$$\frac{p_{1i}}{p_{0i}} = \frac{P(\gamma_1 = i\,|\,\gamma_0 = 1)}{P(\gamma_1 = i\,|\,\gamma_0 = 0)} \quad \text{(A.3)}$$

are the BFs in favor of the consultand being a carrier, given the 1° relative's carrier status. Since carrier status is unknown, the BF is a weighted average over the possible status, with weights M_{ij} reflecting how likely the 1° relative is a carrier.

The transmission probabilities $p_{ij} = P(\gamma_1 = j \,|\, \gamma_0 = i)$ to the 1° relative depend on the mode of transmission. To illustrate these calculations, let the mode be autosomal dominant and let π be $P(\gamma_i = 1)$. If γ_1 is a child, then by averaging over the unknown unrelated spouse of the consultand denoted γ_p (i.e., $P(\gamma_p \,|\, \gamma_0) = P(\gamma_p)$),

$$\begin{aligned} p_{01} &= P(\gamma_1 = 1 \,|\, \gamma_p = 0, \gamma_0 = 0) P(\gamma_p = 0 \,|\, \gamma_0 = 0) \\ &\quad + P(\gamma_1 = 1 \,|\, \gamma_p = 1, \gamma_0 = 0) P(\gamma_p = 1 \,|\, \gamma_0 = 0) \\ &= 0 + \frac{\pi}{2} = \frac{\pi}{2} \\ p_{11} &= P(\gamma_1 = 1 \,|\, \gamma_p = 0, \gamma_0 = 1) P(\gamma_p = 0 \,|\, \gamma_0 = 1) \\ &\quad + P(\gamma_1 = 1 \,|\, \gamma_p = 1, \gamma_0 = 1) P(\gamma_p = 1 \,|\, \gamma_0 = 1) \\ &= \frac{1-\pi}{2} + \pi = \frac{1+\pi}{2}. \end{aligned}$$

If γ_1 is a parent, then by averaging over their unknown unrelated spouse,

$$p_{01} = \frac{P(\gamma_1 = 1)}{P(\gamma_0 = 0)} P(\gamma_0 = 0 \,|\, \gamma_1 = 1) = \frac{\pi}{1-\pi} \times \frac{1-\pi}{2} = \frac{\pi}{2}$$

$$p_{11} = \frac{P(\gamma_1 = 1)}{P(\gamma_0 = 1)} P(\gamma_0 = 1 \,|\, \gamma_1 = 1) = \frac{\pi}{\pi} \times \frac{1+\pi}{2} = \frac{1+\pi}{2}.$$

$$\mathrm{BF} = \mathrm{BF_c} \times \frac{\dfrac{1-\pi}{2-\pi} + (1+\pi^{-1}) \left(\dfrac{f_1}{f_0}\right)^{c_1} \left(\dfrac{S_1}{S_0}\right)^{1-c_1} \dfrac{\pi}{2} \times \dfrac{f_0^{c_2} S_0^{1-c_2}(1-\pi) + f_1^{c_2} S_1^{1-c_2}(1+\pi)}{f_0^{c_2} S_0^{1-c_2}(2-\pi) + f_1^{c_2} S_1^{1-c_2} \pi}}{1 + \left(\dfrac{f_1}{f_0}\right)^{c_1} \left(\dfrac{S_1}{S_0}\right)^{1-c_1} \dfrac{\pi}{2} \times \dfrac{f_0^{c_2} S_0^{1-c_2}(1-\pi) + f_1^{c_2} S_1^{1-c_2}(1+\pi)}{f_0^{c_2} S_0^{1-c_2}(2-\pi) + f_1^{c_2} S_1^{1-c_2} \pi}}.$$

The p_{ij} are the same for either type of 1° relative. The BFs of (A.3) are

$$\frac{p_{10}}{p_{00}} = \frac{1-\pi}{2-\pi}, \quad \frac{p_{11}}{p_{01}} = 1 + \frac{1}{\pi}. \tag{A.4}$$

Under a rare autosomal-dominant mutation, which assumes that the BFs of equation (A.4) are $1/2$ and $1/\pi$, respectively, and that π is rare enough so that $\pi f_1/f_0 \approx 0$ and $\pi S_1/S_0 \approx 0$, plugging (A.4) into (A.2) the BF is

$$\begin{aligned} \mathrm{BF} &= \mathrm{BF_c} \times \frac{\dfrac{1-\pi}{2-\pi} + \left(1 + \dfrac{1}{\pi}\right) \dfrac{f_1^{c_1} S_1^{1-c_1}}{f_0^{c_1} S_0^{1-c_1}} \dfrac{\pi}{2}}{1 + \dfrac{f_1^{c_1} S_1^{1-c_1}}{f_0^{c_1} S_0^{1-c_1}} \dfrac{\pi}{2}} \\ &\approx \mathrm{BF_c} \times \frac{1}{2} \left(1 + \frac{f_1^{c_1} S_1^{1-c_1}}{f_0^{c_1} S_0^{1-c_1}}\right). \end{aligned} \tag{A.5}$$

Consultand/1° Relative/2° Relative Family

Here the 2° relative is directly related to the 1° relative, for example, mother and maternal grandmother. Assuming conditional independence of phenotypes given a member's carrier status, the likelihood is

$$\begin{aligned} P(H \,|\, \gamma_0) &= P(H_0 \,|\, \gamma_0) \\ &\quad \times \sum_{\gamma_1, \gamma_2 = 0}^{\gamma_1, \gamma_2 = 1} P(H_1 \,|\, \gamma_1) P(H_2 \,|\, \gamma_2) P(\gamma_2 \,|\, \gamma_1) P(\gamma_1 \,|\, \gamma_0), \end{aligned}$$

and the BF is

$$\begin{aligned} \mathrm{BF}(H) &= \frac{P(H \,|\, \gamma_0 = 1)}{P(H \,|\, \gamma_0 = 0)} = \mathrm{BF_c} \times \frac{M_{10} G M_0 + M_{11} G M_1}{M_{00} G M_0 + M_{01} G M_1} \\ &= \mathrm{BF_c} \times \frac{\dfrac{p_{10}}{p_{00}} \dfrac{1}{M_{01} G M_1} + \dfrac{p_{11}}{p_{01}} \dfrac{1}{M_{00} G M_0}}{\dfrac{1}{M_{01} G M_1} + \dfrac{1}{M_{00} G M_0}}, \end{aligned} \tag{A.6}$$

where $GM_i = P(H_2 \,|\, \gamma_1 = i)$. The family's contribution is a weighted average of the BFs if the 1° relative's carrier status is known. To simplify cumbersome expressions, whenever f_1/f_0 or S_1/S_0 is raised to the c_i power, evaluate those ratios at y_i. For an autosomal-dominant mutation, the BF is

Under a rare autosomal-dominant mutation approximation, the BFs of equation (A.4) are $1/2$ and $1/\pi$, respectively, and π is rare enough so $\pi(f_1/f_0)^2 \approx 0$, $\pi(S_1/S_0)^2 \approx 0$, and $\pi(f_1/f_0)(S_1/S_0) \approx 0$. Then the BF is approximately

$$\mathrm{BF} \approx \mathrm{BF_c} \times \frac{1}{2} \left\{ 1 + \frac{f_1^{c_1} S_1^{1-c_1}}{f_0^{c_1} S_0^{1-c_1}} \times \frac{1}{2} \left(1 + \frac{f_1^{c_2} S_1^{1-c_2}}{f_0^{c_2} S_0^{1-c_2}}\right) \right\}. \tag{A.7}$$

The recursive structure suggests how higher-degree relatives are included.